z/VM Resource Manager

WAVV Conference
2004

Bill Bitner
bitnerb@us.ibm.com
IBM Endicott

Based on presentation by
Christine Casey

Borrowed from Chuck Morse

General Resource Control

- Primarily addresses individual machines
  - Dedicating resources
  - Favoring access to resources
- Various controls exist
  - Initial settings via directory statements
  - Other changes via CP commands
- Resources
  - Processor
  - Storage
  - I/O

VM Resource Manager Objectives

- Manage workloads to CPU and DASD I/O velocity goals
- Allow I/O priority queuing to be exploited on behalf of VM-based workloads
- Provide an infrastructure for more extensive workload management for future releases of z/VM

VM Resource Manager Overview

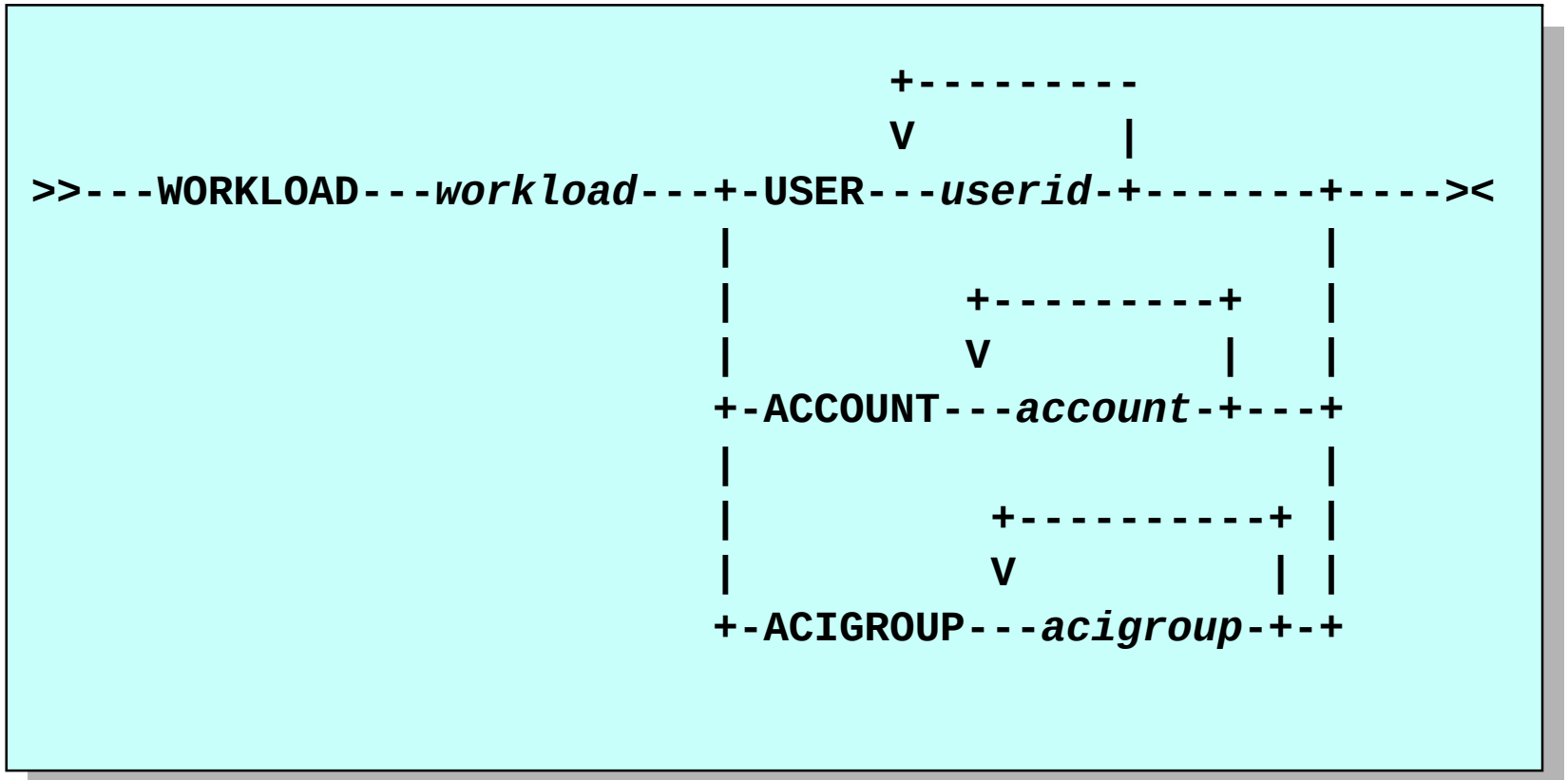
- VM Resource Manager (VMRM) new in z/VM 4.3.0 May 2002
- New service virtual machine - VMRMSVM
  - Uses VM monitor data to obtain regular measurements of virtual machine resource consumption.
  - Based on a customer-supplied definition of workloads, goals and priorities, adjusts virtual machine tuning parameters to achieve those goals.
  - Approximately once a minute, the SVM
    - Computes the achievement levels of interest for each workload
    - For each goal type (DASD and CPU)
      - Selects one workload to adjust based on the customer-supplied importance value

VMRM CONFIG File

- The PROFILE EXEC for VMRMSVM begins operation of the server by calling the IRMSERV EXEC
- IRMSERV reads the customer-supplied definition file
  - Default is VMRM CONFIG A
  - Any other file name can be passed to the IRMSERV EXEC
- The VMRM CONFIG file supports 4 types of statements:
  - WORKLOAD - describes a workload by userid, account, acigroup
  - GOAL - describes a DASD or CPU velocity goal
  - MANAGE - associates a workload with a goal
  - ADMIN - identifies a user to receive VMRM server messages
- The server will not start if there are ANY errors in the VMRM CONFIG file

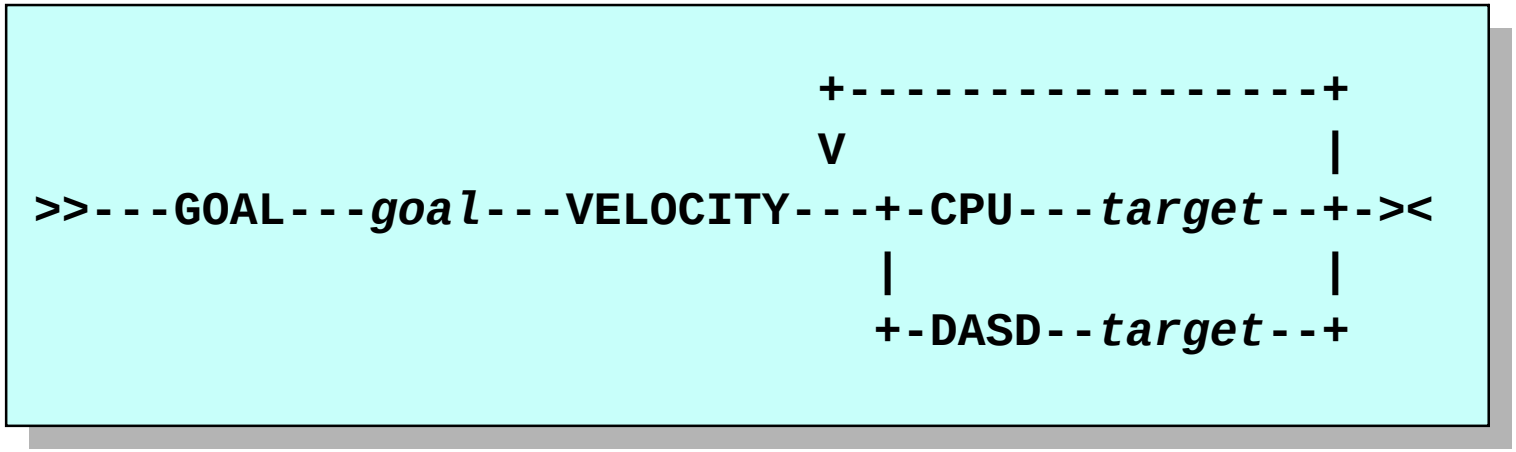
WORKLOAD Statement

- A workload is comprised of one or more virtual machines identified by user ID, account, or ACI group name



GOAL Statement

- The GOAL statement specifies velocity goals for
  - DASD - percentage of time that the user's DASD I/O requests are not outprioritized
  - CPU - percentage of the time the user should receive CPU resources when it is ready



- The CPU and DASD operands can only be specified once
- No adjustments are made if the user is within 5% of goal

MANAGE Statement

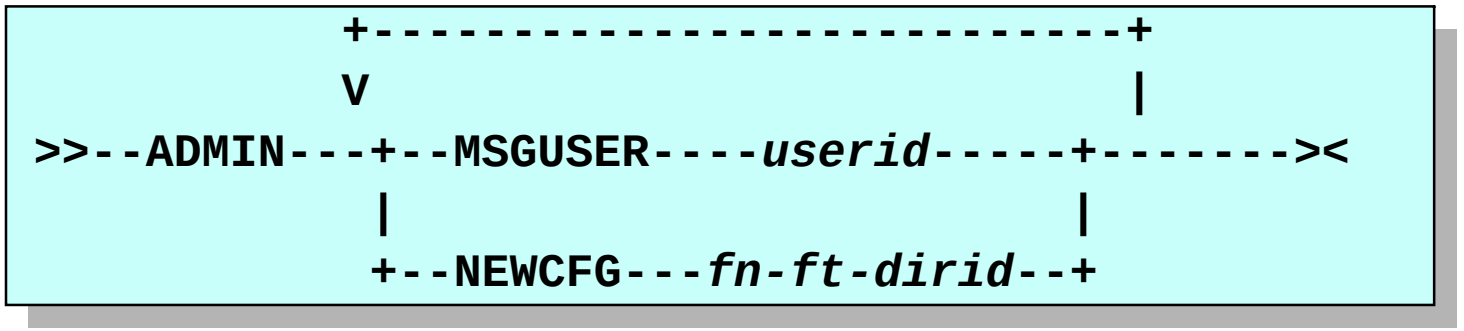
- Associates a workload with a goal
- Assigns an importance value to the relationship
  - Importance values can range from 1-10 (10 is most important)

```
>>---MANAGE---workload---GOAL---goal---IMPORTANCE---value--->< 
```

Only one manage statement is allowed for each workload

ADMIN Statement

- Specifies a user ID on the same system where messages can be sent from the service virtual machine if necessary
  - Messages will be logged to **VMRM LOG1 A**, whether or not there is a user Id specified to receive them
  - Additional service machine events will also be logged in that file
  - **VMRM LOG1 A** will be copied to **VMRM LOG2 A** when it reaches 10,000 records. **VMRM LOG1** will then be erased and rewritten.



If multiple ADMIN statements exist, only the last will be used

NEWCFG added in z/VM 4.4.0

Other z/VM 4.4.0 Enhancements

- Performance improvements
- Messages and debug information
- Syntax checking for config file
  - IRMSERV *config_file* (SYNCHECK
- Dynamically update config file when located on SFS directory and included in ADMIN statement of config file
  - ADMIN *fn ft dirid*
- Wildcards in Workload statements
- Monitor APPLDATA

Sample VMRM CONFIG File

```
*   This is a valid comment line  *
/*  So is this                    */
;   and this
ADMIN    MSGUSER  Chris
WORKLOAD work1    USER abcde a123 456
WORKLOAD work2    USER fghij
WORKLOAD workabcd USER qrst
WORKLOAD work3    ACCOUNT 1234 5678
WORKLOAD work4    ACIGROUP  ABC
GOAL   goal1 VELOCITY CPU  10
GOAL   goal2 VELOCITY DASD 50
GOAL   goal3 VELOCITY CPU  80  DASD 20
MANAGE work1 GOAL goal1 IMPORTANCE 10
MANAGE work2 GOAL goal1 IMPORTANCE  5
MANAGE work3 GOAL goal2 IMPORTANCE  2
MANAGE work4 GOAL goal3 IMPORTANCE 10
MANAGE workabcd   GOAL  goal2 IMPORTANCE  7
```

Sample VMRM log file

```
Each statement has a timestamp similar to the following:

2002-02-19 17:02:02 ServExe  MSG

MSG      IRMSER0022I VM Resource Manager Initialization started
PCfg     VMRM CONFIG A1 2/19/02 17:01:55
MSG      IRMSER0008W The ADMIN message user ID is not logged on..
InitEnv  Monitor sample started -- recording is pending
InitEnv  HCPMNR6224I Sample recording is pending because there...
InitEnv  MONITOR EVENT INACTIVE   BLOCK    4   PARTITION    0
InitEnv  MONITOR DCSS NAME - NO DCSS NAME DEFINED
InitEnv  CONFIGURATION SIZE        68 LIMIT      1  MINUTES
InitEnv  CONFIGURATION AREA IS FREE
InitEnv  USERS CONNECTED TO *MONITOR - NO USERS CONNECTED
InitEnv  .
InitEnv  more data from Q Monitor...
InitEnv  .
MSG      IRMSER0023I VM Resource Manager Initialization complete.
         Proceeding to connect to Monitor.
Exit     STARMON completed. RC=0
ExitSVM  Monitor sample stopped
MSG      IRMSER0012I VM Resource Manager shutdown in progress
```

Workload Selection

- Selection criteria
  - Workloads are selected first based on their workload importance value
  - If a workload was selected in the last interval, either for improvement or degradation, it is skipped and an attempt is made to select another
  - If there are workloads of equal importance, the workload farthest from its goal is selected
  - All users within a workload will have their SHARE or IOPRIORITY adjusted appropriately based on how far they are from the workload goal

Processor Share Terminology

- Absolute vs. Relative Share
  - **Absolute** specifies a user is to receive a target minimum of nnn% of the scheduled system resources
  - Amount of resources available to relative share users = total resources available less the amount allocated to absolute share users
  - **Relative** portion that the user receives is nnnn / sum of all relative share users
  - VM Resource Manager will **not** adjust Absolute users

- Limithard vs. Limitsoft
  - **Limithard** specifies the user's share of CPU resource is limited (they do not receive more than maximum share of the CPU resource)
  - **Limitsoft** specifies that the user's share of CPU resource is limited, **but** the limit can be exceeded if the capacity is available

Adjustment Algorithms

- Individual users within the selected workload may be adjusted based on calculations from monitor data
- For CPU goals:
  - User must have a Relative SHARE setting
  - User does not have Limithard specified on their CPU SHARE setting
  - Sum of wait deltas and run deltas is > current sample size of 5
  - CPU actual = run delta / (run delta + wait delta) * 100
- For DASD goals:
  - User must have a Relative I/O Priority setting
  - Sum of I/O deltas and Outprioritized deltas is > current sample size of 5 for DASD
  - DASD actual = IO delta / (IO delta + outprior delta) * 100
- After above criteria is met, if user is not within 5% of workload goal, then they can be adjusted.

Adjustment Algorithms

- How much to adjust each user?
  - For CPU goals:
    relvalue = (Workload CPU goal / User actual) * User current share

    ** checking that value falls within 1-10,000 range

  - For DASD goals:
    relvalueLo = (Workload DASD goal / User actual) * User curr IO Lo

    relvalueHi = relvalueLo + (User curr Hi - User curr Lo)

    **checking that values fall within 0-255 range

- Set Share and/or Set IOPriority command is issued on behalf of the user

I/O Priority Queuing

- Enables prioritization of virtual machine I/O
  - Allows a guest's I/O priority queuing range
    - To be set via
      - IOPRIORITY directory statement
      - SET IOPRIORITY command
    - To be queried via QUERY IOPRIORITY command
  - If I/O Priority Queuing is available and enabled (zSeries only)
    - I/O Priority Queuing low/high range is obtained from the hardware
    - Guest I/O Priority Queuing values are mapped to fall within that range
    - CP I/O uses highest value available
  - If not available or enabled
    - CP simulates a range of 0-255
  - For I/O priority-aware guests, the priority associated with the guest I/O requests will be enforced
  - For non I/O priority-aware guests, CP assigns a priority value

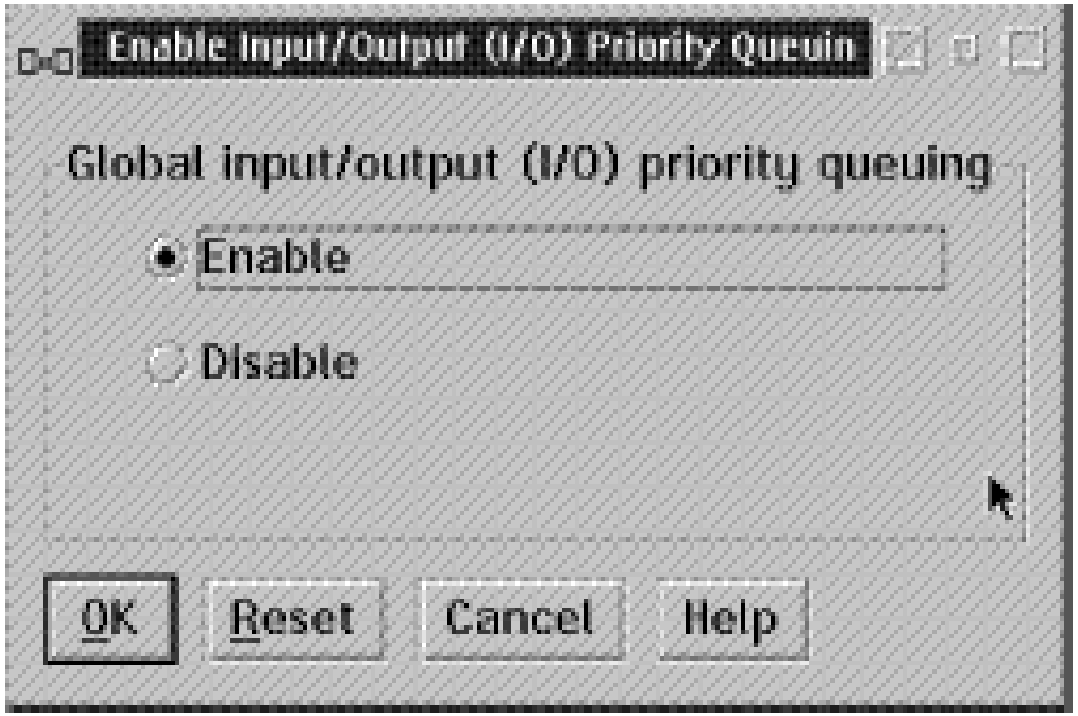
I/O Priority Queuing Mappings

- Mapping of requested range to "effective" range is based on whether hardware facility exists:

	Relative	**Absolute**
Hardware Not Enabled	0 - 255 on command maps to simulated effective range of 0 - 255	0 -255 on command maps to simulated effective range of 0 - 255
Hardware Enabled	0 - 255 on command maps proportionally to hardware range	User input maps directly to hardware range

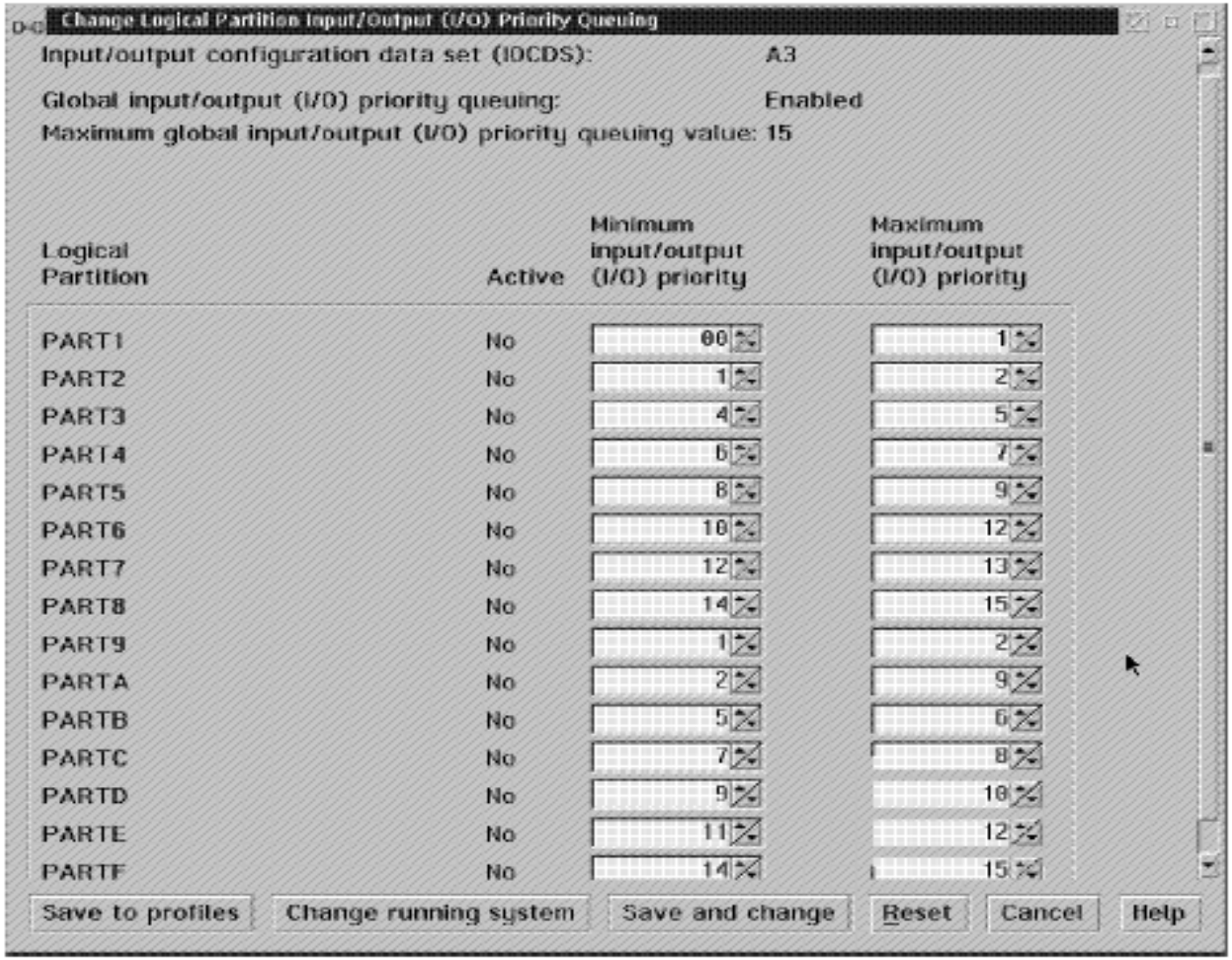
Enabling I/O Priority Queuing on zSeries Processors

- At the Hardware Management Console
  - Use the Enable I/O Priority Queuing task available from the Central Processor Complex Operational Customization tasks list to either enable or disable I/O priority queuing for the entire CPC



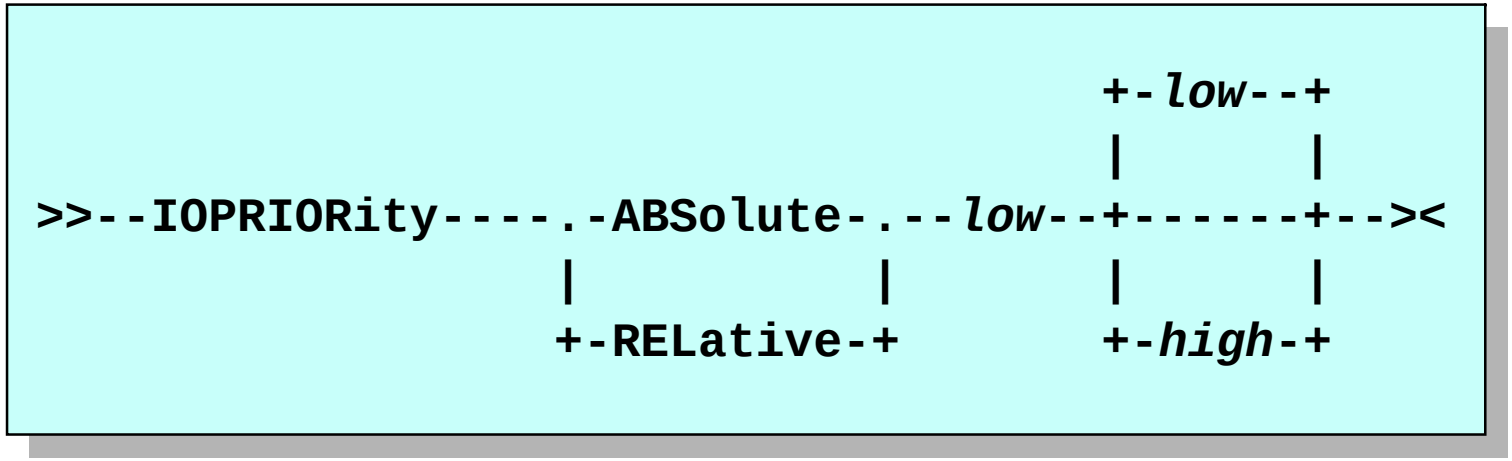
HW I/O Priority Queuing Ranges

- Use the change LPAR I/O priority queuing task to set the minimum and maximum I/O priority queuing values



IOPRIORITY Directory Statement

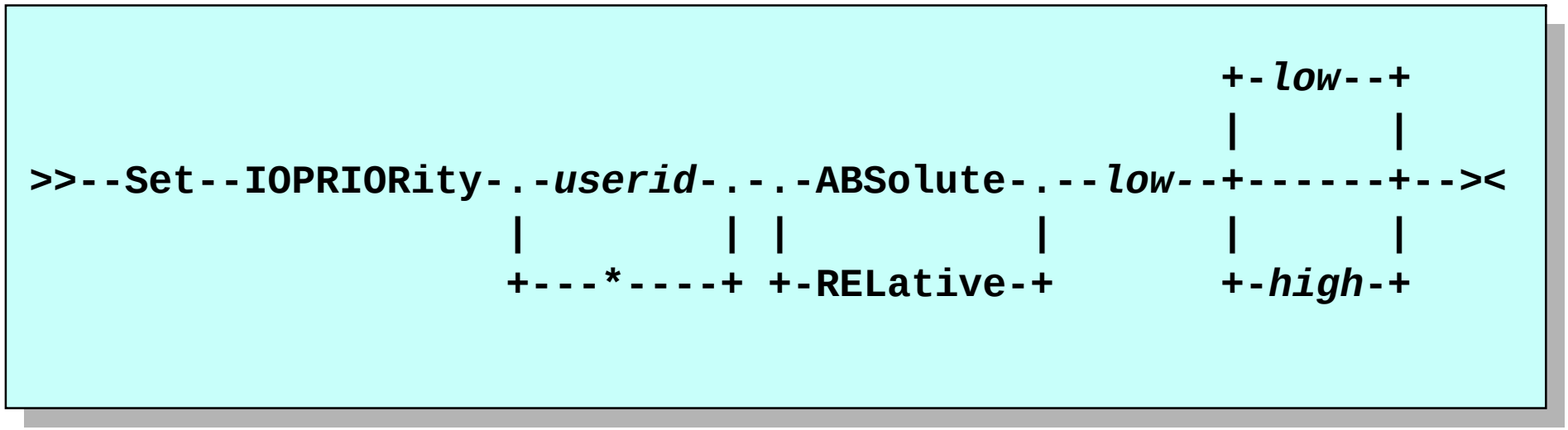
- Specifies the I/O priority range to be set when the user logs on
  - Low and high values must be decimal numbers from 0 to 255
  - If hardware priority queuing is available and enabled
    - Absolute priority ranges outside the range available to CP are clipped to fall within that range
    - Relative ranges are mapped to fall within the range available to CP



If IOPRIORITY is not specified in the directory, low and high are set to a relative value of 0

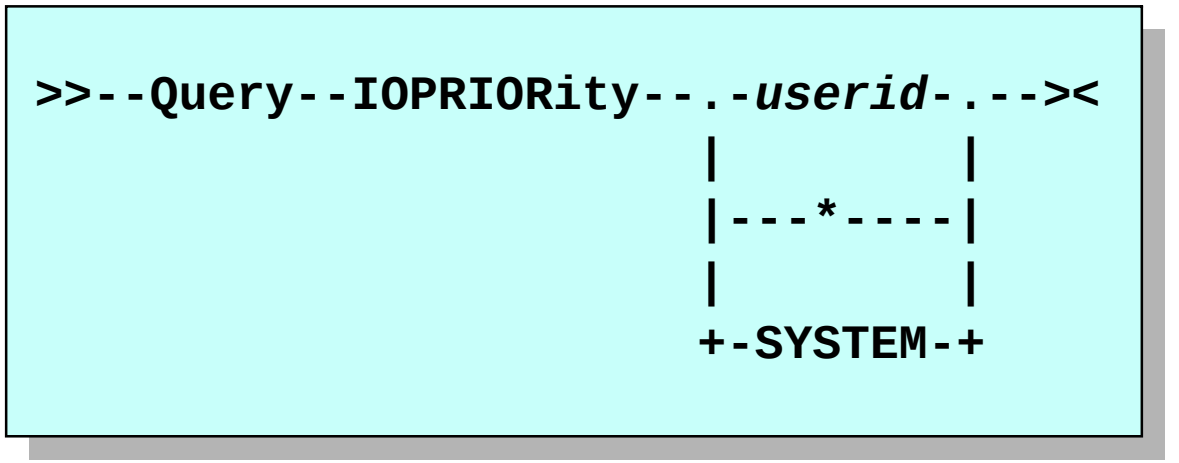
SET IOPRIORITY

- A class A user can adjust a guest's I/O Priority Queuing range



QUERY IOPRIORITY

- A class A or E user can display a guest's or the system I/O Priority Queuing range



  - **userid** requests the priority range of a given userid
  - ***** requests the priority range of the user issuing the command
  - **SYSTEM** requests the priority range available to CP

Query IOPRIORITY Responses

- userid REQUESTED RANGE nnn mmm ABSOLUTE
  EFFECTIVE RANGE xxx yyy

- userid REQUESTED RANGE nnn mmm RELATIVE
  EFFECTIVE RANGE xxx yyy

where:

requested range indicates low and high ranges requested for this user

effective range is the low and high range that CP will allow for this user

Examples of Absolute I/O Priority Queuing Ranges

- If the I/O priority queuing range available to CP is 50-75
  - Virtual machine requests for ranges from 0-49 will assigned an absolute value of 50
  - Virtual machine requests for ranges 50-75 will be accepted
  - Virtual machine requests for ranges 75-255 will be assigned an absolute value of 75

Examples of Relative I/O Priority Queuing Ranges

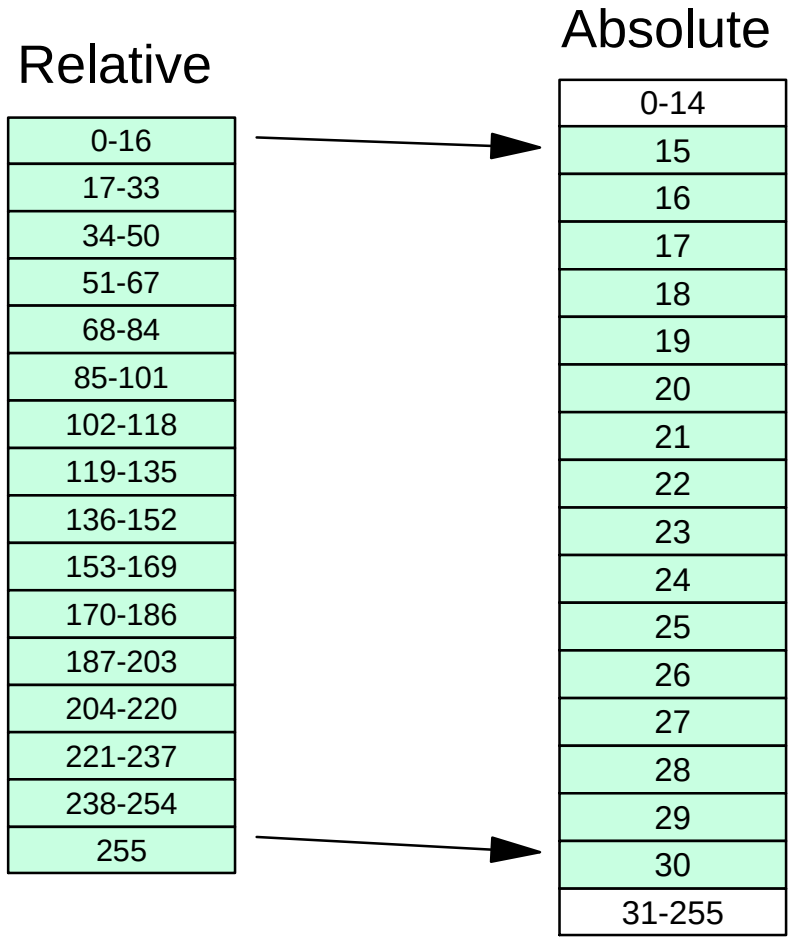
- The effective I/O priority queuing value is calculated from the requested value and the range available to CP

$$Eff\_Val = TRUNC\left(\frac{Rel\_Val * (CP\_Hi - CP\_Lo)}{255}\right) + CP\_Lo$$

- Where:
  - **Eff_Val** is the effective I/O priority
  - **Rel_Val** is the relative I/O priority
  - **CP_Hi** is the highest I/O priority value available to CP
  - **CP_Lo** is the lowest I/O priority value available to CP

Examples of Relative I/O Priority Queuing Ranges

- If the range of I/O priority values available to CP is 15-30 then relative priorities map to absolute priorities as follows:



New Monitor Data

- Existing Monitor records updated
  - User Domain - User Activity Data - D4R3
    - Relative or absolute I/O priority
    - requested and effective priority range
    - Number of times DASD I/O requests have been outprioritized
  - System Domain - User Data - D0R8
    - I/O Priority Queuing Active flag
    - High & low values available to CP
- New Monitor record
  - Scheduler Domain - I/O Priority Queuing Changes - D2R11
    - Event record when I/O priority queuing values change for a user
      - SET IOPRIORITY command
      - Range available to CP changes

Summary

- VMRM provides a method to manage large numbers of virtual machines
  - a flavor of group scheduling
- Resource areas
  - processor over commitment
  - queuing on DASD devices
- See Performance Manual for details
- Monitor data showing goals and actual statistics for workloads reported by Performance Toolkit

- Customer requirements ... we welcome your feedback!
  - Other workload goals you wish to see managed ?